Zaawansowana analityka biznesowa
- siła modeli predykcyjnych

Temat: Praktyczne zastosowanie modeli oceny wartości relacji z klientem w czasie:
MODEL SEGMENTACJI

Filip Krysztaszek

Kamil Książek

Natalia Sadownik

Marianna Gładysz

Piotr Żołnierczyk

Alan Kashkash

1. Wprowadzenie

Współcześnie firmy przechowują i przetwarzają ogromne ilości informacji w bazach i hurtowniach danych. Zgromadzone dane opisujące działania przedsiębiorstwa jak i interakcji z klientami pozwalają między innymi na analizę trendów, anomalii rozwoju firmy, oceny zachowań i preferencji klientów oraz na grupowanie konsumentów.

Grupowanie, inaczej zwane klasteryzacją, pozwala na stworzenie skończonych podzbiorów obiektów posiadających podobne cechy, co umożliwia odpowiednie określenie ich charakteru celem podejmowania racjonalnych decyzji biznesowych firmy np. ukierunkowanie akcji marketingowych. Jednym ze sztandarowych przykładów klasteryzacji jest grupowanie klientów ze względu na podobieństwo zakupionych przez nich produktów lub ilość zrealizowanych transakcji, a skuteczny marketing oparty na analizie tych danych i klasteryzacji pozwala na wyodrębnienie klientów potencjalnie zainteresowanych innymi produktami. Jednym z obszarów wykorzystania, gdzie segmentacja może okazać się bardzo cennym narzędziem jest bankowość, czy też e-commerce. Posiadając dane dotyczące zachowania konsumentów oraz posiadając informacje podawane przez klientów, można wyodrębnić tych, wobec których najlepiej zastosować promocje oraz ukierunkować akcje marketingowe.

1.1. Cel projektu

Celem projektu jest wykonanie segmentacji brazylijskiego e-commerce za pomocą metody k-średnich na podstawie danych udostępnionych na platformie kaggle.com. Klasteryzacja pozwoli wydzielić grupy dla obszernego spisu transakcji i danych o produktach, w których podobieństwo pewnych cech jest maksymalizowane. W projekcie skupiono się na zbadaniu czy dostępne transakcje są możliwe do sklasteryzowania poprzez zastosowanie metody k-średnich.
Projekt został wykonany w języku Python.

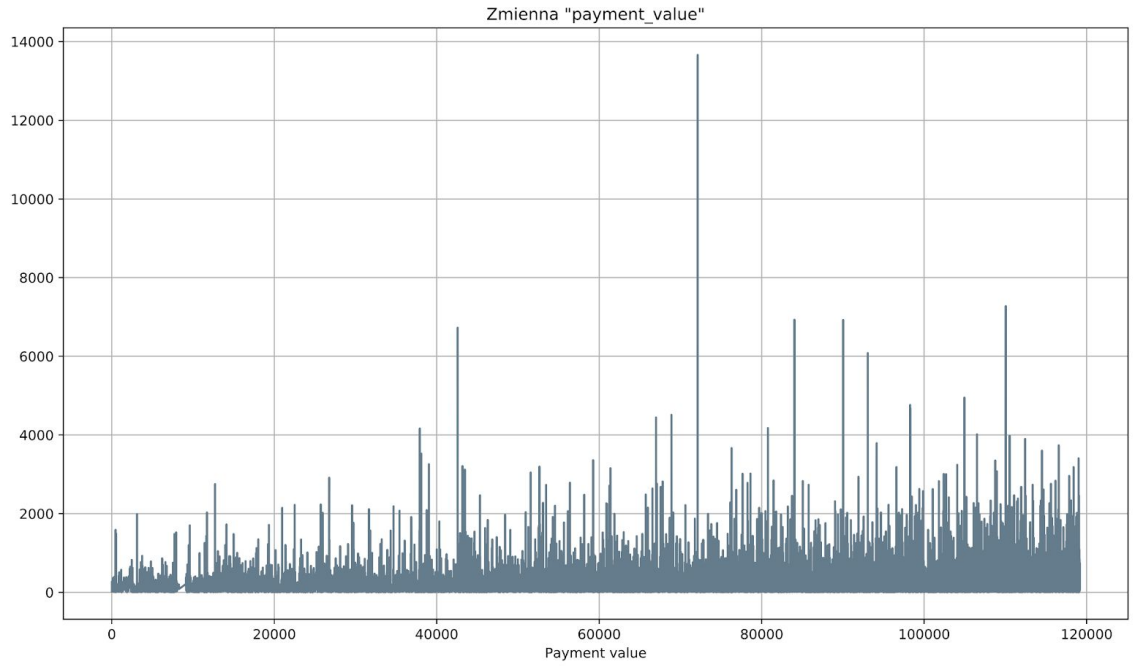
1.2. Opis danych

Dane “Brazilian E-Commerce Public Dataset by Olist” pochodzą ze strony kaggle.com. Jest to zbiór danych z brazylijskiego e-commerce o zamówieniach dokonanych w Olist Store. Zestaw zawiera informacje o 100 tys. zamówień od 2016 do 2018 roku dokonanych na wielu

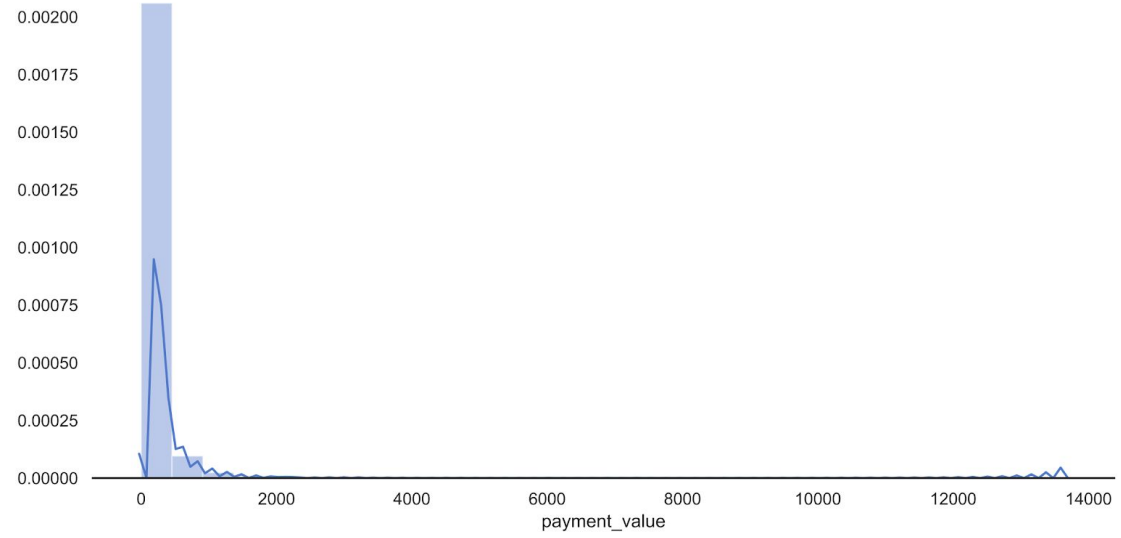
brazylijskich rynkach. Jego funkcje pozwalają na przeglądanie zamówienia w wielu wymiarach: od statusu zamówienia, ceny, płatności i wyników frachtu do lokalizacji klienta, atrybutów produktów i wreszcie opinii napisanych przez klientów. Są to prawdziwe dane handlowe, które zostały zanonimizowane. Zbiór danych został dostarczony przez Olist, największy dom towarowy na brazylijskich rynkach. Olist łączy małe firmy z całej Brazylii jednym kontraktem. Handlowcy ci są w stanie sprzedawać swoje produkty za pośrednictwem Olist Store i wysyłać je bezpośrednio do klientów za pośrednictwem partnerów logistycznych Olist. Więcej informacji można znaleźć na stronie internetowej: www.olist.com

2. Analiza struktury

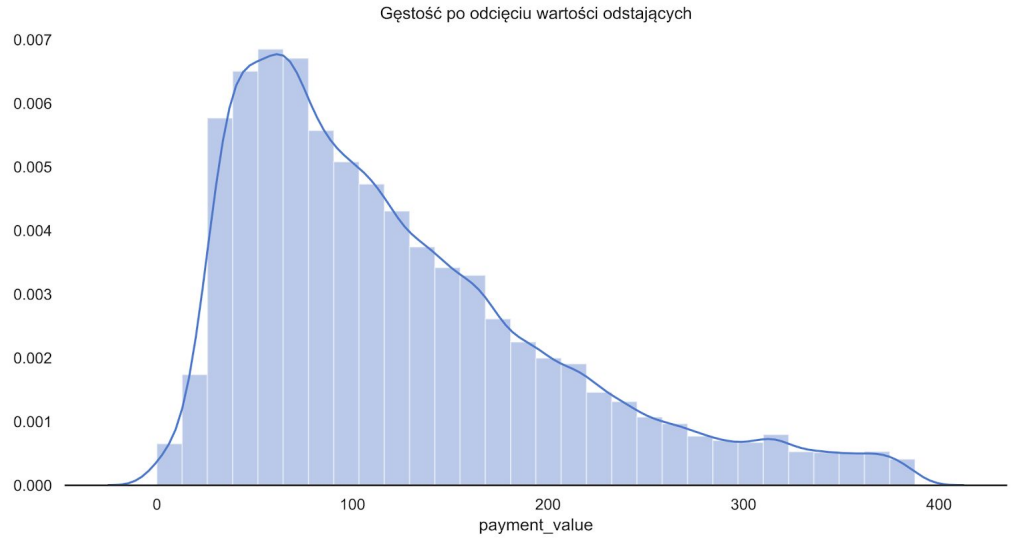
Celem analizy struktury jest przeprowadzenie opisu statystycznego badanej zbiorowości. Opis stworzony został z wykorzystaniem podstawowych miar, które w zwięzły sposób oddają właściwości zbioru. Zbiór danych poddany analizie został przekształcony. Na wstępnej selekcji usunięto zmienne które nie były użyteczne z punktu rozpatrywanej analizy. Zbiór danych został zawężony pod kątem ujednolicenia metody płatności oraz statusu przesyłki. Wybrano tylko te obserwacje które miały wyłącznie jeden sposób płatności oraz tylko te których zamówienia zakończyły się uregulowaną płatnością i dostarczeniem przesyłki do klienta. Następnie zbiór danych został sprawdzony pod względem braków danych. Usunięto 1542 obserwacji. Zmienna payment value została dokładnie zbadana, sprawdzono jej rozkład oraz wartości odstające które później usunięto.

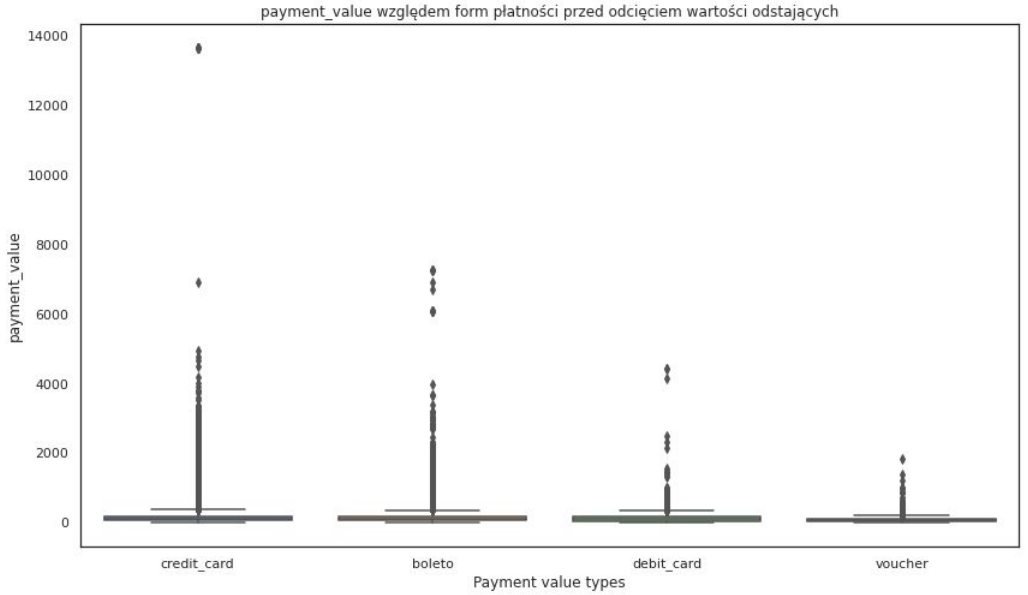


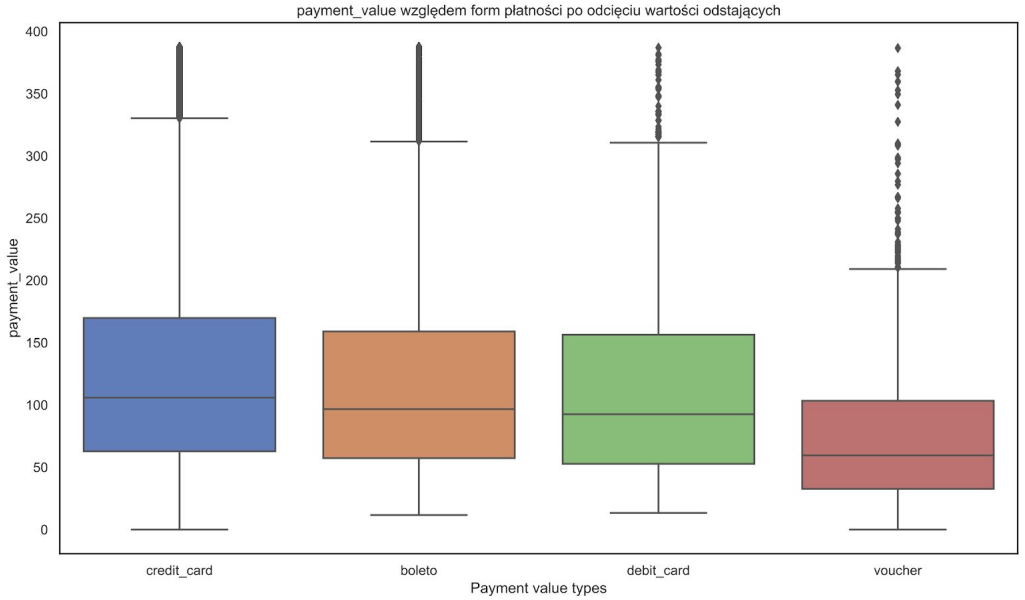
Gęstość rozkładu zmiennej Payment Value przed usunięciem wartości odstających



Gęstość rozkładu zmiennej Payment Value po usunięciu wartości odstających

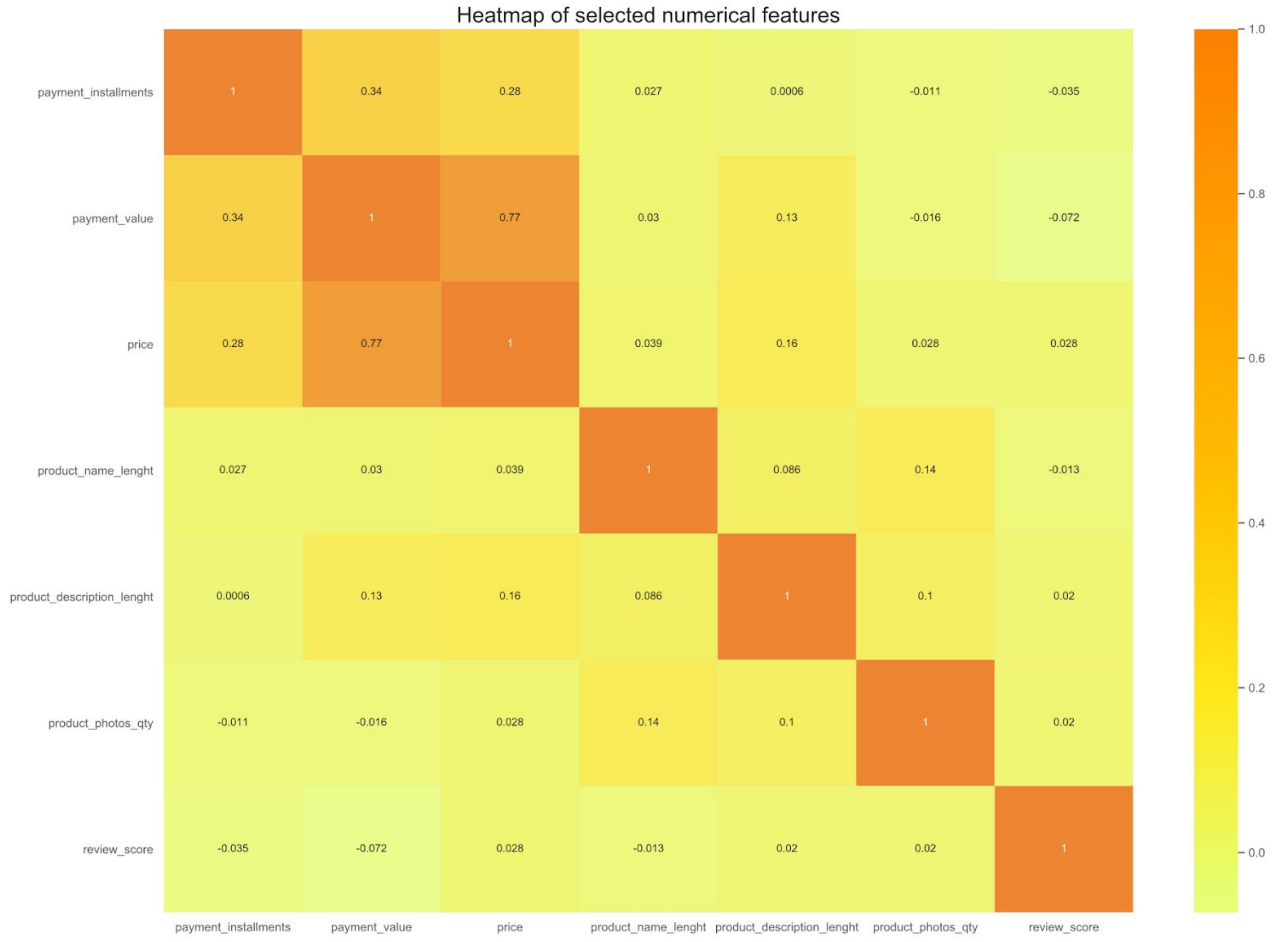

Gęstość po odcięciu wartości odstających
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
100
200
300
400
payment_value


payment_value względem form płatności przed odcięciem wartości odstających
14000
12000
10000
8000
6000
4000
2000
0
payment_value
credit_card
boleto
debit_card
voucher
Payment value types


payment_value względem form płatności po odcięciu wartości odstających
400
350
300
250
200
150
100
50
0
payment_value
credit_card
boleto
debit_card
voucher
Payment value types

3. Klasteryzacja danych

W celu wykonania segmentacji klientów, postanowiono wykorzystać algorytm uczenia nienadzorowanego K-means, służący do iteracyjnej klasteryzacji danych. Aby wyodrębnić najlepsze zmienne do użycia w algorytmie, wykorzystano macierz korelacji, którą następnie przedstawiono na mapie ciepła :



Największe korelacje dodatnie zauważono pomiędzy zmiennymi:

1. payment_value & price
2. payment_installments & payment_value
3. payment_installments & price

Największe korelacje ujemne zauważono pomiędzy zmiennymi:

1. payment value & review_score
2. payment_installments & review_score
3. payment_value & product_photos_qty

Najbardziej skorelowane ze sobą zmiennie zostały odrzucone z dalszego rozważania, gdyż nie niosą istotnej wartości informacyjnej. Postanowiono skupić się na znalezieniu mniejszej korelacji, które występują pomiędzy zmiennymi oraz wyborze zmiennych ciągłych. Zdecydowano o utworzeniu nowej zmiennej uwzględniającej wskaźnik czasu (‘month’) w celu znalezienia dodatkowych zależności i możliwym zapobiegnięciu występowania powiązań pomiędzy już wybranymi nisko skorelowanymi zmiennymi.

Wybrane zmienne w klasteryzacji

Pierwsze badanie klasteryzacji - użyte zmienne:

- product_description_length
- payment_value

Drugie badanie klasteryzacji - użyte zmienne:

- payment_value,
- product_description_lenght,
- month (zmienna month utworzona na podstawie order_purchase_timestamp)

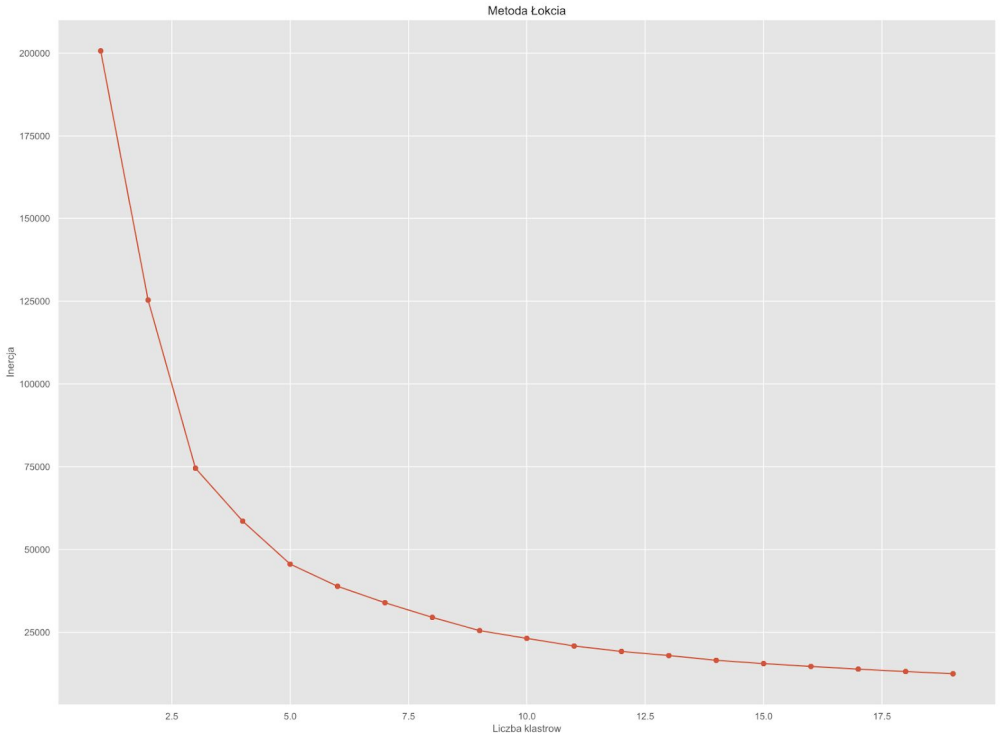
Metoda klasteryzacji

W projekcie uwzględniono także wielokrotne rozwiązania klastrowe z zastosowaniem różnych ilości ziaren początkowych, w celu zagwarantowania zbieżności z optimum globalnym. W języku Python istnieje zaimplementowany rozbudowany algorytm dla metody k-means w pakiecie sci-kit learn o nazwie “k-means++”, którego użycie zapobiega utknięciu algorytmu w pierwotnej fazie inicjacyjnej i lokalnym minimum, następuje wielokrotne losowe rozłożenie położenia centroidów w przestrzeni względem określonej wybranej ilości klastrów, dążąc do lepszego oddalenia od siebie centroidów. K-means jest w tym podobny do działania algorytmu EM Expectation-Maximization, po kilkukrotnych próbach iteracyjnych algorytm wybiera najlepsze położenie centroidów z przeprowadzonych dotychczas rozwiązań uwzględniając równocześnie wartości cechujące się najniższa sumą kwadratów odległości obiektów wewnątrz klastra.

Klasteryzacja 1

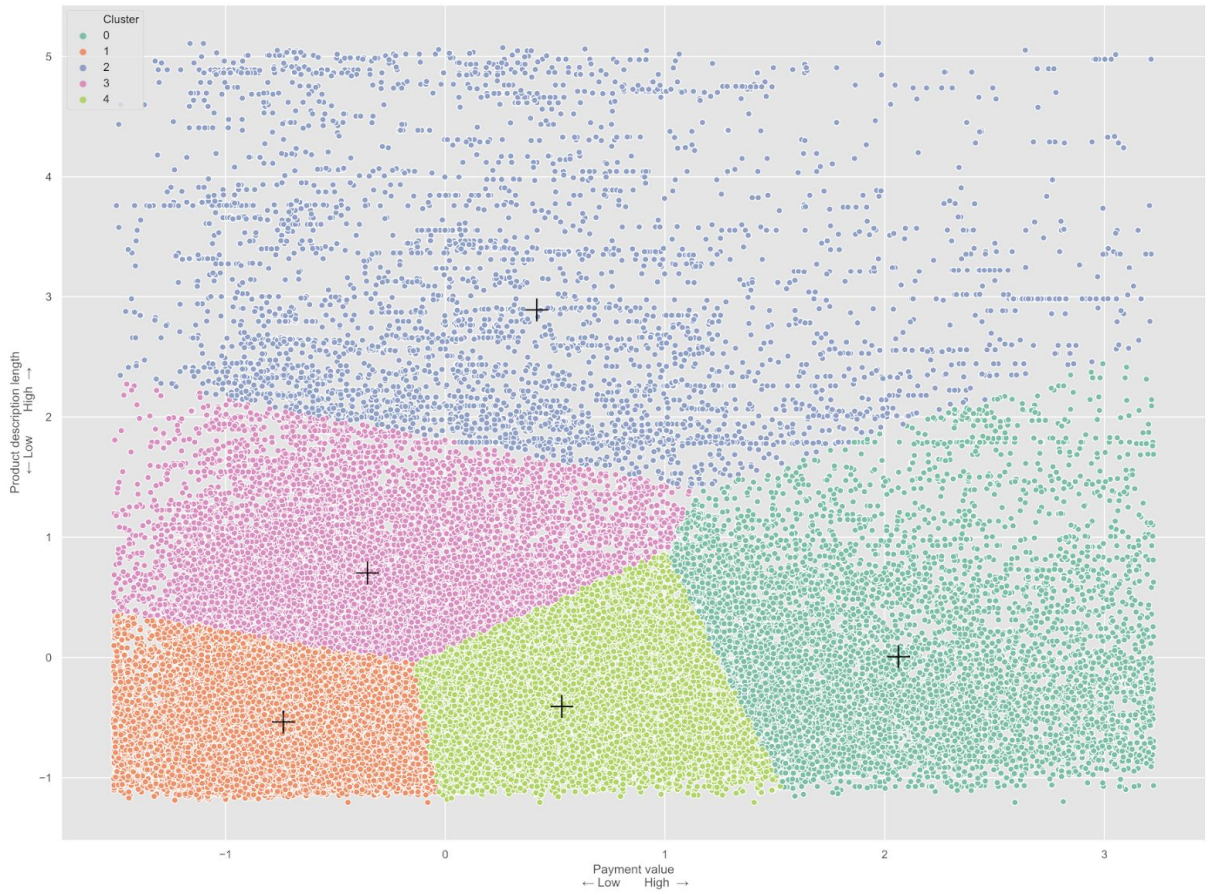
W projekcie przewidziano również oszacowanie optymalnej liczby klastrów użytych przy wyliczeniu modelu k-średnich. Metoda osypiska lub tzw. metoda łokcia, dzięki oszacowaniu wartości inercji z każdego przeprowadzonego modelowania względem ilości klastrów, możliwy jest wybór optymalnej liczby klastrów powyżej której różnice w wariancji są znikome.

Na podstawie Metody Łokcia dla klasteryzacji pierwszej wybrane zostało 5 klastrów.



Na poniższym wykresie można zauważyć skupiska (klastry) utworzone względem zrealizowanych transakcji z klientami w zależności od długości opisu produktów oraz wartości zamówienia. Liczba otrzymanych klastrów:

Klaster	**Ilość obserwacji**
2	41113
5	23289
4	18997
1	10684
3	6261



Klaster 1 - transakcje klientów którzy wybierali produkty z długim opisem i mieli różne wartości zamówień - jest ich znacznie mniej niż innych klientów - ten segment klientów moglibyśmy wyodrębnić w kampaniach marketingowych np. wyświetlając im dodatkowy opis produktu na stronie, wysyłając maile w których szczegółowo opisujemy produkty.

Klaster 2 - transakcje klientów którzy wybierali produkty z krótkim opisem i mieli niskie bądź średnie wartości zamówień

Klaster 3 - transakcje klientów którzy wybierali produkty z krótkim opisem i mieli niskie wartości zamówień

Klaster 4 - transakcje klientów którzy wybierali produkty z krótkim opisem i mieli średnie wartości zamówień

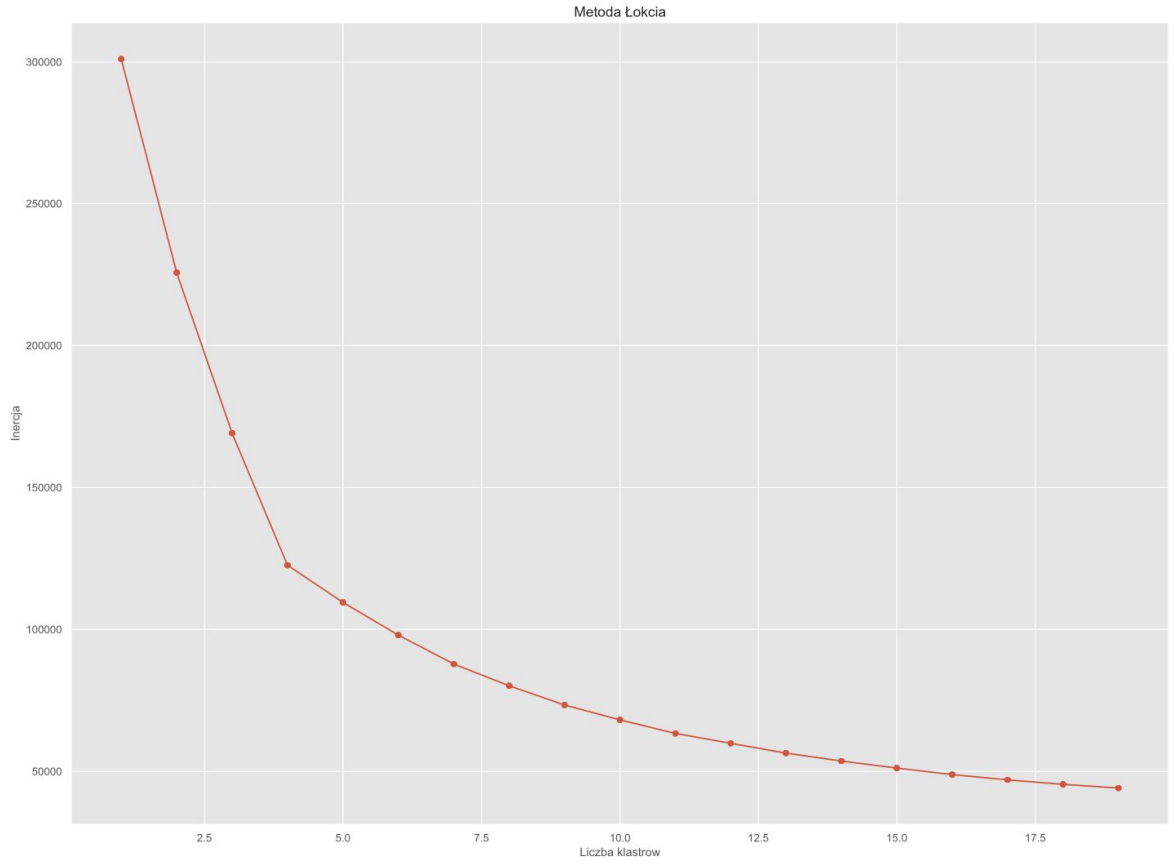
Klaster 5 - transakcje klientów którzy wybierali produkty z krótkim opisem i mieli wysokie wartości zamówień

Klienci z klastrów 2, 3, 4, 5 są to osoby które wybierały produkty z krótkim opisem - Sposób wyświetlania produktów na stronie powinien uwzględniać zwarty opis, najlepiej aby cechy

były wypunktowane. Maile powinny mieć ograniczoną ilość treści, np zdjęcie produktu, nazwa i cena. Ze względu na filtry spamu, pod obrazami umieszczony powinien być tekst alternatywny w postaci szczegółowego opisu produktu - dostawcy usług mailowych nie zablokują takich maili (mała ilość tekstu w mailu = spam) a w przypadku problemów z wyświetlaniem obrazów klienci zobaczą alternatywny opis.

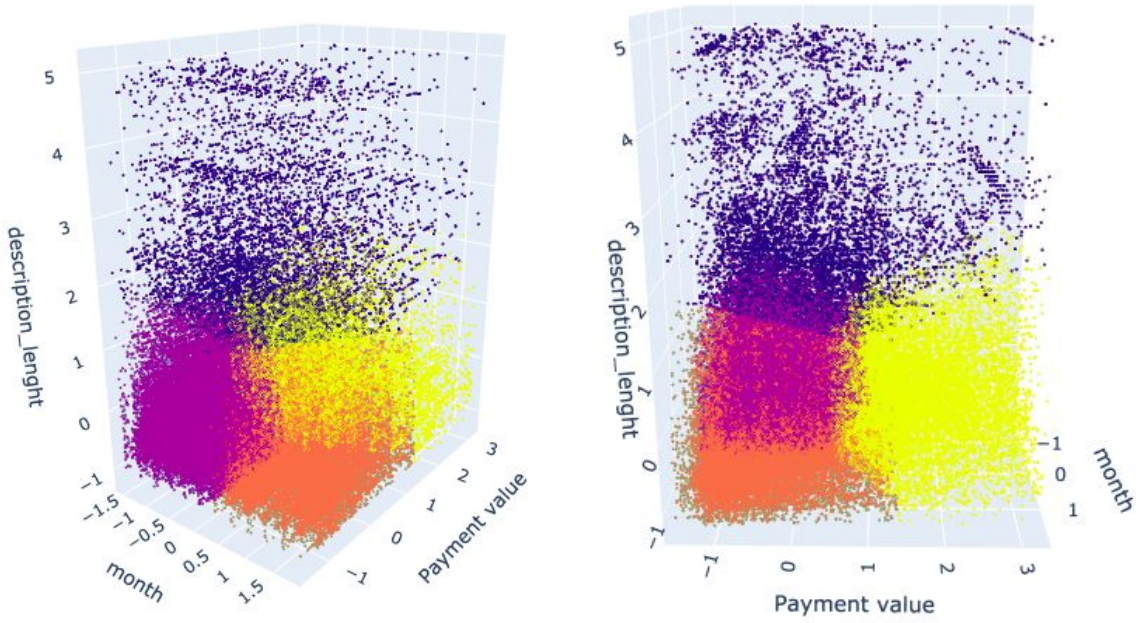
Klasteryzacja 2

Na podstawie Metody Łokcia dla drugiej klasteryzacji ilość wybranych segmentów wynosi 4.



Dodatkowa zmienna “month” przyjęta do klasteryzacji 2 została wybrana dodatkowo aby zbadać sezonowość zakupów.

Klaster	**Ilość obserwacji**
1	39991
4	32991
2	18165
3	9197



Klaster 1 - transakcje cechujące się krótkim opisem, małą wartością zamówienia oraz zostały zamówione w pierwszej połowie roku;

Klaster 2 - podobny do klastra 1: transakcje cechujące się krótkim opisem i małą wartością zamówienia, ale zostały zamówione w drugiej połowie roku;

Klaster 3 - transakcje obejmujące produkty z krótkim lub średnim opisem, cechujące się dużą wartością zamówienia. Dotyczy całego roku; oraz

Klaster 4 - wszystkie transakcje z długim opisem

Można zauważyć utworzenie pewnych grup przy sezonowości. Tak postawa jedna kategoria zamówień niezależna od sezonowości, oraz dwie wyodrębnione w zależności od pierwszej lub drugiej połowy roku. Dla zakupów produktów z długim opisem można zobaczyć nieznaczne zmniejszenie gęstości w pierwszym kwartale roku.

4. Podsumowanie i wnioski

Celem projektu było dokonanie analizy zbioru zawierającego dane z brazylijskiego e-commerce i dokonanie segmentacji za pomocą metody k-średnich, która przyniesie wartość biznesową. Po zbadaniu dostępnych informacji i analizie struktury postanowiono skupić się na zmiennej payment_value, która określa wartość dokonanych transakcji.

Zmienna payment_value, pomimo że nie najlepiej skorelowana z innymi zmiennymi, miała największy potencjał w odkryciu klastrów mogących przynieść wartość biznesową. Oprócz

niej w celu dokonania segmentacji postanowiono uwzględnić zmienne product_description_length oraz month. Po użyciu Metody Łokcia wywnioskowano, że optymalna ilość grup wynosi: 5 dla klasteryzacji obejmującej wartość transakcji i długość opisu oraz 4 na klasteryzacji trójwymiarowej uwzględniającej również miesiąc zakupu.

Klasteryzacja pierwsza wydzieliła grupy pomiędzy którymi zarysowuje się silny podział ze względu na długość opisu produktu. Aż 4 grupy cechuje krótki bądź średniej długości opis produktu, podczas gdy te z dłuższym nie cieszą się dużą popularnością, niezależnie od wartości transakcji. Prawdopodobnie wynika to ze niechęci klientów do czytania bardzo długiego tekstu, gdy starają się wybrać produkt z dostępnych na rynku. Zaleca się wysłanie informacji do sprzedawców z klastra 1 z długim opisem, by rozważyli jego skrócenie oraz poinformowanie sprzedawców z pozostałych klastrów, że długość opisu ich produktu ma istotne znaczenie dla osiągnięcia jak największej ilości dokonanych transakcji.

Klasteryzacja druga ponownie ukazała istotę problemu długich opisów produktów oraz wyodrębniła pewne grupy pod kątem sezonowości. Niezależnie od pory roku pewne produkty cieszą się niezmienną popularnością, a niektóre mają popyt tylko w określonych miesiącach. Podobnie jak w przypadku poprzedniej segmentacji zaleca się poinformowanie ich sprzedawców o zaistniałej sytuacji oraz być może wykluczenie ofert niektórych produktów w pierwszej połowie roku dla których istnieje popyt tylko w drugiej połowie roku. Również warto rozważyć wysłanie informacji do pozostałych oraz zaimplementowanie automatycznego ostrzeżenia, gdy opis produktu jest za duży. Takie rozwiązania pozwolą na zwiększenie zakupów w internecie poprzez brak zniechęcania klientów przez produkty z długim opisem oraz większą konkurencyjność produktów na rynku, ponieważ te wcześniej omijane zaczną być częściej kupowane.

5. Najważniejsze części kodu

 # Wybór zmiennych do modelu

```
fig = plt.figure()
plt.rcParams['figure.figsize'] = (20, 15)
plt.style.use('ggplot')

sns.heatmap(data.corr(), annot = True, cmap = 'Wistia')
plt.title('Heatmap of selected numerical features', fontsize = 20)
```

```
plt.show()
fig.savefig("heatmap.pdf")
```

 # Clustering 1

```
 from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
data_good = df.copy()
data_good.iloc[:,:-1] = scaler.fit_transform(data_good.iloc[:,:-1])
df_good = pd.DataFrame(data_good,
columns=['payment_value','product_description_lenght','month'])
df_good.head()
from sklearn.cluster import KMeans
wcss = []
for i in range(1, 20):
   kmeans = KMeans(n_jobs = -1, n_clusters=i, init='k-means++',
max_iter=100, n_init=10, random_state=0)
   kmeans.fit(df_good.iloc[:,:-1])
   wcss.append(kmeans.inertia_)

fig = plt.figure()
plt.plot(range(1, 20), wcss, marker='8')
plt.title('Metoda Łokcia')
plt.xlabel('Liczba klastrow')
plt.ylabel('Inercja')
plt.show()
fig.savefig("Inercja_C1.pdf")
```

 # Clustering 2

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler(
data_good2 = df.copy()
data_good2 = scaler.fit_transform(data_good2)
df_good2 = pd.DataFrame(data_good2,
columns=['payment_value','product_description_lenght','month'])
df_good2.head()
wcss = []
for i in range(1, 20):
   kmeans = KMeans(n_jobs = -1, n_clusters=i, init='k-means++',
max_iter=100, n_init=10, random_state=0)
   kmeans.fit(data_good2)
   wcss.append(kmeans.inertia_)
```

```
fig = plt.figure()
plt.plot(range(1, 20), wcss, marker='8')
plt.title('Metoda Łokcia')
plt.xlabel('Liczba klastrow')
plt.ylabel('Inercja')
plt.show()
fig.savefig("Inercja_C2.pdf")
```

6. Informacje o grupie

Lp	Imię i Nazwisko	nr albumu	email	zadania
1	Natalia Sadownik (os kontaktowa)	102691	ns102691@student.sgh.waw.pl	Transformacja zmiennych przed budową modelu, wnioski badawcze
2	Marianna Gładysz	102343	mg102343@student.sgh.waw.pl	Rozwiązania klastrowe oraz wybór najlepszego na podstawie uzyskanych wyników
3	Piotr Żołnierczyk	102382	pz102382@student.sgh.waw.pl	Transformacja zmiennych przed budową modelu, wnioski badawcze
4	Alan Kashkash	72088	ak72088@student.sgh.waw.pl	Wygenerowanie wielokrotne rozwiązań klastrowych dla różnych ziaren początkowych
5	Filip Krysztaszek	84572	fk84572@student.sgh.waw.pl	Kmeans++. Szczegółowa interpretacja wyników dla najlepszego modelu segmentacji + przedstawienie wizualne klastrów. Wnioski badawcze.
6	Kamil Książek	80573	kk80573@student.sgh.waw.pl	Szczegółowa interpretacja wyników dla najlepszego modelu segmentacji + przedstawienie wizualne klastrów